Missing Data Assignment

Name: Vidita Vijaykumar Daga
Student Id: 014488630
Group members: Garima Chapekar, Hussein Adams, Danquing Zhao

Part1: Below are the replacement strategies that are experimented:

1) Replace by mean of all non-null age values

Replacing by mean is in general a reasonable strategy since the mean reflects the average value in the dataset. In this, the dataset properties like mean, standard deviation, and median changed but still, the difference isn't very significant.

2) Replace by mean of all non-null age values after removing outliers (with a standard deviation of 2)

Replacing the mean is a good starting point but we can try to exclude outliers. This strategy removes the outliers that are outside two standard deviations from the mean. (In theory, 2 standard deviations account for about 95% of data). With this approach, while the mean and standard deviation are similar to strategy 1, there is an improvement in percentile properties.

3) Random sampling from Normal Distribution (mean and standard deviation from input)

We can take a different approach and not replace every value with the same value. To do this, if we assume that the age values follow normal distribution curve, we can create a curve and replace the null age values with random values from this curve. With this approach, it is observed that the resultant dataset properties like standard deviation and percentile are even closer to the original values.

4) Random sampling from Normal Distribution excluding outliers

The above strategy with normal sampling computes mean and standard deviation from the input data. It can have outliers and cause noise in the dataset. Hence, to further improve the random sampling based on the normal distribution, we can remove outliers before doing the sampling. While this approach showed very less difference as compared to strategy 3, it still gave better results in dataset properties like the percentile values (25%, 50%, 75%).

5) Replace by average age after grouping by sex

This strategy again considers mean to replace the null age values but it goes to a finer level by computing the average after diving the dataset into 2 groups based on sex. We compute the average age of all males and replace the null age values of all males by that value. Similarly, the average age of all females is used to replace the female null age values. This strategy might not give an accurate result but we can evaluate this pattern to check if this relationship exists in the dataset. After evaluating, this strategy did not give close results as compared to the other strategies

Conclusion for Part1 :

Based on the observations, the 4th strategy: *Random sampling on a normal distribution excluding outliers* gave the closest results and worked best for this dataset. The properties like mean, standard deviation, percentile values (25%, 50%, 75%) were closer to the original dataset thus preserving the original set to a good extent.

Part 2:

For this part, I grouped the dataset by *sex* and *survived* to check if any specific group is affected more by the replacement strategy. The replacement strategy used for this exercise is the *"Random Sampling from Normal Distribution excluding outliers".*

In both cases, the result doesn't affect any specific group differently. The reason is that since we are performing random sampling based on the normal distribution,

the strategy doesn't consider a person's sex or survival factor to determine the value to be replaced. So, the strategy is not biassed towards any specific group.

Dataset properties output of all the strategies (Please check Jupyter Notebook for more details)

Original set:

count	714.000000
mean	29.699118
std	14.526497
min	0.420000
25%	20.125000
50%	28.000000
75%	38.000000
max	80.000000

Strategy 1: Mean

count	891.000000
mean	29.699118
std	13.002015
min	0.420000
25%	22.000000
50%	29.699118
75%	35.000000
max	80.000000

Strategy 2: Mean without outliers

count	891.000000
mean	29.423716
std	13.013789
min	0.420000
25%	22.000000
50%	28.312774

75% 35.000000
max 80.000000

Strategy 3: Random Sampling on Normal Distribution

count 891.000000
mean 29.789994
std 14.528705
min 0.420000
25% 20.231778
50% 28.684471
75% 38.909055
max 80.000000

Strategy 4: Random Sampling on Normal Distribution excluding outliers

count 891.000000
mean 29.502788
std 14.296668
min 0.420000
25% 20.000000
50% 28.000000
75% 38.000000
max 80.000000

Strategy 5: Mean after grouping by sex

count 891.000000
mean 29.736034
std 13.014897
min 0.420000
25% 22.000000
50% 30.000000
75% 35.000000

max 80.000000